A Comparison Between a kNN based Aproach and a PNN Algorithm for a Multi-Label Classification Problem

Elias Oliveira
Department of Information Science

Patrick Marques Ciarelli
Deptartment of Electrical Engineering

Claudine Gonçalves
Department of Computer Science

Universidade Federal do Espírito Santo
Campus de Goiabeiras, Av. Fernando Ferrari, s/n, Cx Postal 5011, 29060-970 – Brazil.
{elias, pciarelli, claudine}@lcad.inf.ufes.br

Abstract

Techniques for categorization and clustering, range from support vector machines, neural networks to Bayesian inference and algebraic methods. The k-Nearest Neighbor Algorithm (kNN) is a popular example of the latter class of these algorithms. Recently, a slight modification of it has been proposed so that the Multi-Label k-Nearest Neighbor Algorithm (Ml-kNN) can deal better with multi-label classification problems. In this paper we are interested in automatic text categorization, which are becoming more and more important as the amount of text in electronic format grows and the access to it becomes more necessary and widespread. We proposed a Probabilistic Neural Network Algorithm (PNN) tailored to also deal with multi-label classification problems, and compared it against the Ml-kNN algorithm. Our implementation surpass the Ml-kNN algorithm in four metrics typically used in the literature for multi-label categorization problems.

Keywords:Text classification, Machine Learning, Business Activities Classification.

1 Introduction

Automatic text classification and clustering are still very challenging computational problems to the information retrieval (IR) communities both in academic and industrial contexts. Currently, a great effort of work on IR, one can find in the literature, is focused on classification and clustering of generics content of text documents. However, there are still many other important applications to which little attention has hitherto been paid, which are as well very difficult to deal with. One example of these applications is the classification of companies based on their economic activities description, also called mission statements, which represent the business context of the companies' activities, in other words, the business economic activities from free text description by the company's founders.

The categorization of companies according to their economic activities constitute a very important step towards building tools for obtaining information for performing statistical analysis of the economic activities within a city or country. With this goal, the Brazilian government is creating a centralized digital library with the business economic activity descriptions of all companies in the country. This library will serve the three government levels: Federal; the 27 States; and more than 5.000 Brazilian counties. We estimate that the data related to nearly 1.5 million companies will have to be processed every year into more than 1.000 possible different activities. It is important to highlight that the large number of possible categories makes this problem particularly complex when compared

with others presented in the literature [7].

In this paper we proposed a slightly modified version of the standard structure of the probabilistic neural network (PNN) [8] so that we could deal with the multi-label problem faced in this work. We have chosen the PNN classifier because of its implementation simplicity and high computacional speed in the training stage, when compared to others algorithms such as SVM and Backpropagation Neural Networks. The complexity of SVM, for example, grows quadratically with the size of the dataset, being thus a bottleneck for large dataset problems [9]. We compared our approach against the Ml-kNN [10] through our business economic activity descriptions dataset and the PNN showed to be far superior than the Ml-kNN.

This work is organized as follows. In Section 2, we detail more the characteristics of the problem and its importance for the government institutions in Brazil. We describe our probabilistic neural network algorithm in Section 3. In Section 4, the experimental results are discussed. Finally, we present our conclusions and indicate some future paths for this research in Section 5.

2 The Problem

In many countries, companies must have a contract (*Articles of Incorporation* or *Corporate Charter*, in USA), with the society where they can legally operate. In Brazil, this contract is called a *social contract* and must contain the *statement of purpose* of the company – this statement of purpose describe the *business activities* of the company and must be categorized into a legal business activity by Brazilian government officials. For that, all legal business activities are cataloged using a table called National Classification of Economic Activities – *Classificação Nacional de Atividade Econômicas*, (CNAE) [2].

To perform the categorization, the government officials (at the Federal, State and County levels) must find *the semantic correspondence* between the company economic activities description and one or more entries of the CNAE table. There is a numerical code for each entry of the CNAE table and, in the categorization task, the government official attributes one or more of such codes to the company at hand. This can happen on the foundation of the company or in a change of its social contract, if that modifies its economic activities.

The work of finding the semantic correspondence between the company economic activities description and a set of entries into the CNAE table are both very difficult and labor-intensive task. This is because of the subjectivity of each local government officials who can focus on their own particular interests so that some codes may be assigned to a company, whereas in other regions, similar companies, may have a totally different set of codes. Having inhomogeneous ways of classifying any company everywhere in all the three levels of the governmental administrations can cause a serious distortion on the key information for the long time planning and taxation. Additionally, the continental size of Brazil makes this problem of classification even worse.

In addition, the number of codes assigned by the human specialist to a company can vary greatly, in our dataset we have seen cases where the number of codes varied from 1 up to 109. However, in the set of assigned codes, the first code is the main code of that company. The remaining codes have no order of importance.

For all these reasons, the computational problem addressed by us is mainly that of automatically *suggesting* the human classifier the semantic correspondence between a textual description of the economic activities of a company and one or more items of the CNAE table. Or, depends on the level of certainty the algorithms have on the automatic classification, we may consider bypassing thus the human classifier.

2.1 Evaluating the Results

Typically, text categorization is mainly evaluated by the *Recall* and *Precision* metrics [1]. Nonetheless, the classification problem presented here has many rare classes (see Table 1) and some experiments have shown that Precision and F1 measures may not be adequate metrics for evaluation this kind of problem [5]. Thus we are going to adopt a set of more appropriated metrics for this type of problem [10].

Formalizing the problem we have at hand, text categorization may be defined as the task of assigning documents to a predefined set of categories, or classes [7]. In multi-label text categoriza-

tion a document may be assigned to one or more categories. Let $\mathcal{D}$ be the domain of documents, $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ a set of pre-defined categories, and $\Omega = \{d_1, d_2, \ldots, d_{|\Omega|}\}$ an initial corpus of documents previously categorized by some human specialists into subsets of categories of $\mathcal{C}$.

In multi-label learning, the training(-and-validation) set $TV = \{d_1, d_2, \ldots, d_{|TV|}\}$ is composed of a number documents, each associated with a subset of categories in $\mathcal{C}$. TV is used to train and validate. Actually, to tune eventual parameters of categorization systems that associate the characteristics of each document in the TV to the appropriate combination of categories. The test set $Te = \{d_{|TV|+1}, d_{|TV|+2}, \ldots, d_{|\Omega|}\}$, on the other hand, consists of documents for which the categories are unknown to the automatic categorization systems. After being trained, as well as tuned, by the TV, the categorization systems are used to predict the set of categories of each document in Te.

A multi-label categorization system typically implements a real-valued function of the form $f : \mathcal{D} \times \mathcal{C} \rightarrow \mathbb{R}$ that returns a value for each pair $\langle d_j, c_j \rangle \in \mathcal{D} \times \mathcal{C}$ that, roughly speaking, represents the evidence for the fact that the test document d_j should be categorized under the category $c_j \in C_j$, where $C_j \subset \mathcal{C}$. The real-valued function $f(.,.)$ can be transformed into a ranking function $r(.,.)$, which is an one-to-one mapping onto $\{1, 2, \ldots, |\mathcal{C}|\}$ such that, if $f(d_j, c_1) > f(d_j, c_2)$, then $r(d_j, c_1) < r(d_j, c_2)$. If C_j is the set of proper categories for the test document d_j, then a successful categorization system tends to rank categories in C_j higher than those not in C_j. Additionally, we also use a threshold parameter so that those categories that are ranked above the threshold τ (*i.e.*, $c_k | f(d_j, c_k) \geq \tau$) are the only ones to be assigned to the test document.

We have thus used five multi-label metrics to evaluate the algorithms we are looking into this work. Four of these metrics were discussed in [10].

Tanimoto Distance (tanimoto$_j$) evaluates how many categories, on the total number of categories predicted by the algorithm, is actually part of the right categories assigned by the human specialist.

$$\text{tanimoto}_j = \frac{|\mathcal{P}_j| + |\mathcal{C}_j| - 2|\mathcal{P}_j \cap \mathcal{C}_j|}{|\mathcal{P}_j| + |\mathcal{C}_j| - |\mathcal{P}_j \cap \mathcal{C}_j|}, \qquad (1)$$

where $|\mathcal{C}|$ is the number of categories and $\mathcal{P}_j \cap \mathcal{C}_j$ is the intersection between the set of predicted categories $\mathcal{P}_j$ and the set of appropriate categories $\mathcal{C}_j$ of the test document d_j. The predicted categories are every categories which is higher than the threshold τ.

One-error (one-error$_j$) evaluates if the top ranked category is present in the set of appropriate categories C_j of the test document d_j.

$$\text{one-error}_j = \begin{cases} 0 & \text{if } [\arg\max_{c \in \mathcal{C}} f(d_j, c)] \in C_j \\ 1 & \text{otherwise.} \end{cases} \qquad (2)$$

where $[\arg\max_{c \in \mathcal{C}} f(d_j, c)]$ returns the top ranked category for the test document d_j.

Coverage (coverage$_j$) measures how far we need to go down the rank of categories in order to cover all the possible categories assigned to a test document.

$$\text{coverage}_j = \max_{c \in C_j} r(d_j, c) - 1 \qquad (3)$$

where $\max_{c \in C_j} r(d_j, c)$ returns the maximum rank for the set of appropriate categories of the test document d_j.

Ranking Loss (rloss$_j$) evaluates the fraction of category pairs $\langle c_k, c_l \rangle$, for which $c_k \in C_j$ and $c_l \notin C_j$, that are reversely ordered (*i.e.*, $r(d_j, c_l) < r(d_j, c_k)$) for the test document d_j.

$$\text{rloss}_j = \frac{R_j}{|C_j||\bar{C}_j|} \quad \textit{where} \qquad (4)$$

$$R_j = |\{(c_1, c_2) | f(d_j, c_1) \leq f(d_j, c_2), (c_1, c_2) \in C_j \times \bar{C}_j\}|$$

where $\bar{C}_j$ is the complementary set of C_j in $\mathcal{C}$.

Average Precision (avgprec$_j$)] evaluates the average fraction of categories ranked above a particular category $c \in C_j$ which actually are in C_j.

$$\text{avgprec}_j = \frac{1}{|C_j|} \sum_{c \in C_i} \frac{|\{c' | r(d_j, c') \leq r(d_j, c), c' \in C_j\}|}{r(d_j, c)} \qquad (5)$$

For p test documents, the overall performance is obtained by averaging each metric, that is, tanimoto $= \frac{1}{p} \sum_{j=1}^{p} \text{tanimoto}_j$, one-error $= \frac{1}{p} \sum_{j=1}^{p} \text{one-error}_j$, coverage $= \frac{1}{p} \sum_{j=1}^{p} \text{coverage}_j$, rloss $= \frac{1}{p} \sum_{j=1}^{p} \text{rloss}_j$, and avgprec $= \frac{1}{p} \sum_{j=1}^{p} \text{avgprec}_j$. The smaller the value of *tanimoto distance*, *one-error*, *coverage* and *ranking loss*, and the larger the value of *average precision*, the better the performance of the categorization system. The performance is optimal when tanimoto = one-error = rloss = 0 and avgprec = 1.

3 The Algorithms

The PNN was first proposed by Donald Specht in 1990 [8]. This is an artificial neural network for nonlinear computing which approaches the Bayes optimal decision boundaries. This is done by estimating the *probability density function* of the training dataset using the Parzen nonparametric estimator.

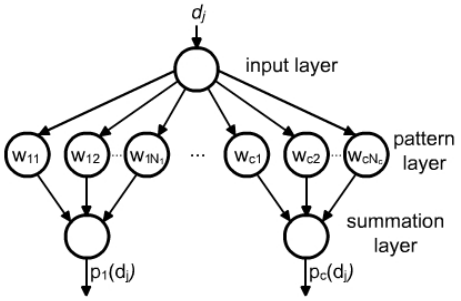

Figure 1: The modified Probabilistic Neural Network architecture.

The literature has shown that this type of neural network can yield similar results, sometimes superior, in pattern recognition problems when compared to the others techniques [4].

The original Probabilistic Neural Network algorithm was designed for uni-label problems. Thus, we slightly modified its standard architecture, so that it is now capable of solving multi-label problems, a type of problems reported in this work.

In our modified version, instead of four, the Probabilistic Neural Network is composed of only three layers: the *input* layer, the *pattern* layer and the *summation* layer, as depicted in Figure 1. Thus like in the original structure, this version of Probabilistic Neural Network needs only one training step, thus its train is very fast comparing to the others feed-forward neural networks [3]. The train consists in assigning each training sample w_j of category C_j to a neuron of pattern layer of category C_j. Thus the weight vector of this neuron is the characteristics vector of the sample.

For each d_j test instance passed by the input layer to a neuron in the pattern layer, it computes the output for the d_j. The computation is as showed in Equation 6.

$$F_{k,i}(d_j) = \frac{1}{2\pi\sigma^2} exp(\frac{d_j^t w_{ki} - 1}{\sigma^2}), \qquad (6)$$

where the d_j is the pattern characteristics input vector, and the w_{ki} is the k^{th} sample for a neuron of category C_i, $k \in N_i$, whereas N_i is the number of neuron of C_i. In addition, d_j was normalized so that $d_j^t d_j = 1$ and $w_{ki}^t w_{ki} = 1$. σ is the Gaussian standard deviation, which determines the receptive field of the Gaussian curve.

The next step is the summation layer. In this layer, all weight vectors are summed, Equation 7, in each cluster C_i producing $p_i(d_j)$ values, where $|C|$ is the total number of categories.

$$p_i(d_j) = \sum_{k=1}^{N_i} F_{k,i}(d_j), \quad i = 1, 2, \ldots, |C| \qquad (7)$$

Finally, for the selection of the categories which will be assigned by neural network to each sample, we consider the most likely categories pointed out by the summation layer based on a chosen threshold.

Differently from other types of networks, such as those feed forward based , the PNN proposed needs few parameters to be configured: the σ, (see in Equation 6) and the determination of threshold value. Other advantages of the probabilistic neural networks is that it is easy to add new categories, or new training inputs, into the already running structure, which is good for the on-line applications [3]. On the other hand, one of its drawbacks is the great number of neurons in the pattern layer, which can be, nevertheless, mitigated by an optimization on the number of the neuron [6].

4 Experiments

In our experiments we chose to compare our approach against the Ml-kNN due to the fact that it is pointed out as yielding the best results on all the different datasets studied in [10]. Therefore, to evaluate the performance of our probabilistic neural network and the Ml-kNN algorithm we used one dataset containing 3264 documents of free text business descriptions of Brazilian companies categorized into a subset of 764 CNAE categories. This

dataset was obtained from real companies placed in Vitoria County in Brazil. The CNAE codes of each company in this dataset were assigned by Brazilian government officials trained in this task. Then we evenly partitioned the whole dataset into four subsets of equal size of 816 documents. We joined to this categorizing dataset the brief description of each one of the 764 CNAE categories, totalizing 4028 documents. Hence, in all training (-and-validation) set, we adopted the 764 descriptions of CNAE categories and a subset of 816 business description documents, and, as the test set, the other three subsets of business descriptions totalizing 2448 documents. As a result, we carried out a sequence of four experiments with each of theses algorithms in order to gather the mean value of their accuracy.

We preprocessed the dataset via term selection –a total of 1001 terms were found in the database after removing stop words and trivial cases of gender and plural; only words appearing in the CNAE table were considered. After that, each document in the dataset was described as a multidimensional vector using the *Bag-of-Words* representation , *i.e.*, each dimension of the vector corresponds to the number of times a term in the 1001 terms vocabulary appears in the corresponding document. Table 1 summarizes the characteristics of this dataset[1].

In Table 1, #C denotes the number of categories, #t denotes the number of terms in the vocabulary, NTD denotes the average number of terms per document, DC denotes the percentage of documents belonging to more than one category, CD denotes the average number of categories of each document, and RC denotes the percentage of rare categories, *i.e.*, those categories associated with less than 1% of the documents of the dataset.

In both PNN and ML-KNN algorithms, their parameters were optimized for each category of the dataset. In the probabilistic neural network case, one value of σ for each category and one value of threshold were selected by a Genetic Algorithm . For the ML-KNN, we also optimized the number of nearest neighbors. To tune these parameters we used the training set, which was used to inductively build the categorizer, and a validation set, which was used to evaluate the performance of the categorizer in the series of experiments aimed at parameter optimization. The training set is composed of 764 descriptions of CNAE categories and the validation set of 816 business description documents described previously.

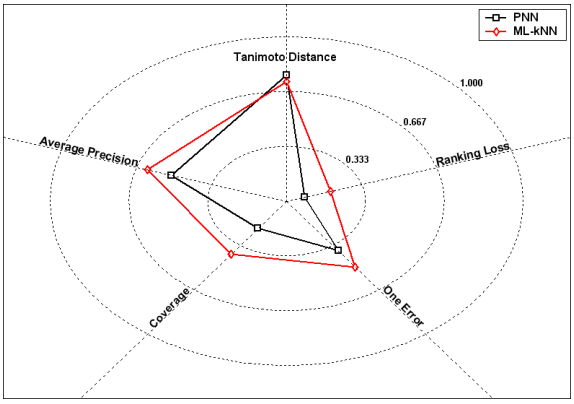

Figure 2: Experimental results of each multi-label categorizer on the economic activities dataset.

After tuning, the multi-label categorizers were trained with the 764 descriptions of CNAE categories of the training set and tested with the 2448 documents of the test set. Figure 2 presents the average experimental results of both multi-label categorization technique: PNN and ML-KNN, on the economic activities dataset in terms of *tanimoto distance*, *ranking loss*, *one-error*, *coverage* and *average precision*, respectively.

In Figure 2, each metric in Figure 2 is represented by a ray, emanating from the center of the circle. Its values varies from 0.0, in the center, to 1.0, on the border of the circle. The result yielded by an algorithm, with respect to a given metric, is then plotted over the appropriated rays. The smaller the value for the *tanimoto distance*, *ranking loss one-error*, and *coverage* metrics, the better. On the other hand, the larger the value for the *average precision*, the better. A normalization on the *coverage* results was devised so that its value could fit between 0 and 1. Therefore, we draw the actual value divided by $|\mathcal{C}| - 1$. Similarly, in order to draw the results of the *average precision* the same way we have done for the other metrics, we are plotting, in Figure 2, the $\overline{\text{average precision}} = 1 - (\textit{average precision})$.

Our approach, as shown by the innermost lines

[1]dataset available at `http://www.inf.ufes.br/~elias/vitoria.tar.gz`.

	#C	#t	Training set				Test/validation set			
			NTD	DC	CD	RC	NTD	DC	CD	RC
CNAE	764	1001	4.65	0.00	1.00	100.00	10.92	74.48	4.27	85.21

Table 1: Characteristics of the CNAE dataset

in Figure 2, outperforms ML-kNN in terms of the four multi-label evaluation metrics adopted, showing differences of 0.1168, 0.1216, 0.1933 and 0.1067 in terms of *ranking loss*, *one error*, *coverage*, and *average precision*, respectively. On the other hand, the PNN algorithm has shown an inferior performance comparing to the ML-kNN on the *tanimoto* metric. The ML-kNN overcome in 0.05 our approach on this metric.

Table 2 shows the numerical values of the results for the compared algorithms.

	PNN	ML-kNN
tanimoto distance	0.7682	**0.7266**
ranking loss	**0.0798**	0.1966
one-error	**0.3736**	0.4952
coverage	**0.2050**	0.3983
$\overline{\text{average precision}}$	**0.5120**	0.6187

Table 2: Numerical results of the comparison of probabilistic neural network and ML-kNN.

where $\overline{\text{average precision}} = 1 - (\textit{average precision})$

5 Conclusions

The problem of classifying huge number of economic activities description in free text format every day is a huge challenge for the Brazilian governmental administration. This problem is crucial for the long term planning in all three levels of the administration in Brazil.

In this work, we presented an experimental evaluation of the performance of Probabilistic Neural Network on multi-label text classification. We performed a comparative study of PNN and the multi-label lazy learning technique ML-kNN [10] using a multi-label dataset for the categorization of free-text descriptions of economic activities. In this problem, PNN outperformed ML-kNN in four from five multi-label evaluation criteria adopted.

A direction for future work is to boldly compare the PNN performance against other multi-label text categorization methods.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, New York, 1 edition, 1998.

[2] CNAE. *Classificação Nacional de Atividades Econômicas Fiscal*. IBGE – Intituto Brasileiro de Geografia e Estatística, Rio de Janeiro, RJ, 1.1 edition, 2003. http://www.ibge.gov.br/concla.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2 edition, 2001.

[4] C. C. Fung, V. Iyer, W. Brown, and K. W. Wong. Comparing the Performance of Different Neural Networks Architectures for the Prediction of Mineral Prospectivity, 2005.

[5] X. Hao, X. Tao, C. Zhang, and Y. Hu. An Effective Method to Improve kNN Text Classifier. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, volume 1, pages 379–384, 2007.

[6] K. Z. Mao, K. C. Tan, and W. Ser. Probabilistic Neural-Network Structure Determination for Pattern Classification. *IEEE Transactions on Neural Networks*, 11:1009–1016, 2000.

[7] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[8] D. F. Specht. Probabilistic Neural Networks. *Neural Networks*, 3(1):109–118, 1990.

[9] J. xiong Dong, A. Krzyzak, and C. Y. Suen. Fast SVM Training Algorithm with Decomposition on Very Large Data Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):603–618, 2005.

[10] M.-L. Zhang and Z.-H. Zhou. ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recogn.*, 40(7):2038–2048, 2007.